Specifying and Implementing Nonparametric and Semiparametric Survival Estimators in Two-Stage (Nested) Cohort Studies With Missing Case Data

Steven D. MARK and Hormuzd A. KATKI

Since 1986, we have been studying a cohort of individuals from a region in China with epidemic rates of gastric cardia cancer and have conducted numerous two-stage studies to assess the association of various exposures with this cancer. Two-stage studies are a commonly used statistical design. Stage one involves observing the outcomes and accessible baseline covariate information on all cohort members, and stage two involves using the stage one observations to select a subset of the cohort for measurements of exposures that are difficult to obtain. When the outcomes are censored failure times, such as in our studies, the most common designs used are the case-cohort and nested case-control designs. One limitation of both these designs is that the estimators of the cumulative hazards, and hence survivals and absolute risks, are biased when some cases are missing the stage two measurements. In our experience, such missingness is present in virtually all two-stage studies that (like ours) use biological specimens to obtain exposure measurements. In earlier work we derived and characterized the efficiency of a class of nonparametric and a class of semiparametric cumulative hazard estimators that are unbiased regardless of whether or not all cases are measured. In this article we limit the presentation of the mathematical derivation of these two classes to aspects important to study design and analysis. We analyze data from a two-stage study that we conducted on the association of *Helicobacter pylori* infection with incident gastric cardia cancers. We discuss the substantive reasons why we deliberately sampled only 25% of the available cancer cases. Through simulations, we demonstrate that substantial variation in precision exists between unbiased estimators within each class, and express the origin of these differences in terms of parameters familiar to investigators. We describe how preexistent knowledge about these parameters can be used to increase estimator precision, and detail specific strategies for constructing such estimators. Computer code in R that implements these estimators is available from the authors on request.

KEY WORDS: Absolute risk; Auxiliary covariate; Case cohort; Case control; Cox proportional hazard; Cumulative hazard; Efficiency; Missing covariates; Nested cohort; Nonparametric; Robust estimation; Semiparametric; Standardized survival; Survival; Two-stage study; Weighted estimating equation.

1. INTRODUCTION

Nearly all large epidemiologic cohort studies initiated in the last 30 years have been designed to estimate the association of a disease with measurements made on biological samples (Samet and Munoz 1998). These measurements, which we denote by V_i and refer to as *exposures*, are typically expensive and consume scarce resources. For example, in the data analysis that we present in Section 6, $V_i = 1$ if an individual has serologic evidence of *Helicobacter pylori* (Hp) infection and $V_i = 0$ otherwise. In an attempt to reduce the number of V_i measurements while minimizing the decrease in the precision of the estimates, numerous sampling designs, called "two-stage studies" by statisticians (Robins, Rotnitzky, and Zhao 1994; hereinafter RRZ), and "nested cohort studies" by epidemiologists (Samet and Munoz 1998), have been proposed. Although the proposals vary with type of data and parameters of interest (see RRZ), the general structure is as follows. At the start of the cohort (time 0), investigators obtain biological specimens as well as measurements on a large number of covariates, A_i. Typically, the A_i is information obtained from such measurement instruments as questionnaires, physical exams, and laboratory tests. Endpoints of interest are recorded up until some time, τ. This constitutes the stage-one data, which we denote by W_i. In stage two, the stage-one data are used to select a subset of individuals on whom V_i will be measured. Understandably, all two-stage designs call for sampling a smaller fraction of controls (i.e., individuals without the observed endpoint) than cases. In fact, it is generally specified that two-stage designs select for V_i measurement "all the cases of interest, but only a subsample of the noncases" (Samet and Munoz 1998, p. 8).

In this article we focus on two-stage studies where the endpoint of interest is censored failure time and the parameters of interest are survival probabilities. The nomenclature for these endpoints, which for concreteness we annotate in terms of our Hp study (Sec. 6), is as follows. T_i is the time to the event of interest, in this case time to the development of gastric cardia cancer (GCC). Instead of T_i, we observed the right-censored event outcome (X_i, Δ_i), where $X_i = \min(T_i, C_i)$, C_i is an independent censoring time, $\Delta_i = I(X_i = T_i)$, and $I(\cdot)$ is the indicator function. As in most large cohorts, the censoring in the Hp study was almost entirely due to the end of follow-up, which in this study was at time, $\tau = 5.25$ years. We refer to individuals with $\Delta_i = 1$ as cases and those with $\Delta_i = 0$ as controls.

A number of two-stage designs have been proposed for estimation with censored failure time endpoints (for a comprehensive review, see Mark and Katki 2001). Nearly all are variations of the original nested-case control (NCC) (Borgan, Goldstein, and Langholz 1995) and case-cohort (CCH) proposals (Prentice 1986; Self and Prentice 1988). The NCC and CCH designs are also by far the most common designs used in practice (Samet and Munoz 1998). The main distinction between the NCC and CCH designs is the control sampling schemes; we briefly review these in Appendix B.

Steven D. Mark is Professor, Department of Preventive Medicine and Biometrics, University of Colorado School of Medicine, Denver, CO 80262 (E-mail: *Steven.Mark@uchsc.edu*). Hormuzd A. Katki is a Staff Scientist, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20852. Mark's work was done while he was a Senior Research Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute. The authors thank Jamie Robins for many helpful discussions, and the editor, associate editor, and several referees for their comments.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Applications and Case Studies
DOI 10.1198/016214505000000952

Both the NCC and CCH designs specify that the risk of failure is related to covariates through a semiparametric Cox proportional hazards model (CPH) such as (1),

$$\lambda(u|Z_i) = \lambda_o(u)\exp(\beta_{o1}^T V_i + \beta_{o2}^T J_i). \quad (1)$$

Here $Z_i = \{V_i, J_i\}$ is a p-dimensional vector of *exposure covariates*, V_i, and *adjusting covariates*, J_i. The adjusting covariates are a subset of all of the baseline covariate information collected at time 0 ($J_i \subseteq A_i$). These are typically covariates, such as age and sex, that investigators want to control for to obtain unconfounded estimates of the association of V_i with T_i. Although the emphasis of both the NCC and CCH designs has been on estimating the parameters $\beta_o = \{\beta_{o1}^T, \beta_{o2}^T\}$, these designs do provide estimators of the cumulative hazards, $\Lambda(t; z)$,

$$\Lambda(t; z) = \int_0^t \lambda(u|z)\,du, \qquad 0 \le t \le \tau; z \in \mathcal{Z}, \quad (2)$$

where $\mathcal{Z}$ is the support of Z_i. Just as estimates of the hazard ratios, which we refer to using the more generic term *relative risks* [$rr(z)$], are obtained from the identity $rr(z) = \exp(\beta_o^T z)$, estimates of survival are obtained from the identity

$$S(t|z) = \exp\{-\Lambda(t; z)\}. \quad (3)$$

An important limitation of the NCC and CCH cumulative hazard estimators is that unlike the estimators of β_o, they are biased if any cases are missing V_i measurements (Mark and Katki 2001; Mark 2003). Mark (2003) derived and characterized the efficiencies of a class of nonparametric and a class of semiparametric cumulative hazard estimators that are unbiased regardless of whether all of the cases are measured. In our two-stage studies there have always been cases that by either chance (see Sec. 6.1) or design (see Sec. 6.2), were missing V_i measurements. Indeed, because of events outside the investigators' control (i.e., *missing by chance*), we suspect that it would be rare for any large cohort study using biological specimens to be able to measure V_i on all cases.

Thus, to obtain unbiased estimates of the effect of V_i on survival in our two-stage studies, we used nonparametric and semiparametric estimators in the classes described by Mark (2003). The primary goals of this article are to describe these classes of unbiased estimators, provide a derivation that displays the determinates of the efficiency of a particular estimator within a class, demonstrate the application of a semiparametric estimator to the data from the Hp study, and, through simulations, both illustrate the wide variation in efficiency that can exist between two unbiased estimators and describe specific strategies to improve the efficiency of an estimator.

The article is organized as follows. In Section 2 we formally state the goals of our inference and the structure of the two-stage studies that we consider. In Section 3 we define the term *full-data estimators*, then, applying the general results of RRZ on obtaining two-stage estimators from full-data estimators, we derive a class of unbiased semiparametric and nonparametric cumulative hazard estimators. For the former, we assume that hazards are specified by a CPH model (1); for the latter, we make no assumptions about hazards at different levels of covariates. In Section 4 we define a general method for implementing our estimators, which we call $\widehat{\pi}$-estimation, and describe how the efficiency of one $\widehat{\pi}$-estimator relates to another. In Section 5 we apply the theorems of RRZ to censored time-to-event data and derive the mathematical form of the most efficient estimator within each class. We reexpress the efficient form in terms of quantities already familiar to researchers involved in observational studies, and show the general implications for study design and analysis. In Section 6 we review features of several of our two-stage studies, and use a $\widehat{\pi}$-estimating procedure to estimate the effect of Hp infection on absolute risks and risk differences. In Section 7 we apply the general formulation of what constitutes efficient estimators presented in Section 5, and propose specific estimators. Simulations demonstrate that the relative efficiencies of these estimators correspond to predictions from theory. Finally, in Section 8 we provide a simple, nontechnical summary of the results and their practical consequences. Annotated code in R (Ihaka and Gentleman 1996) that implements the $\widehat{\pi}$-estimating procedures is available from the authors.

In this article we restrict the mathematical results presented by Mark (2003) to that subset necessary for understanding the practical implications of those results. In addition, to make the presentation more accessible, in the body of the article we express results only in terms of those functionals of the random variables required to understand the proofs and their importance to applications. Actually defining these functionals requires counting process and martingale notation. With the exception of two expressions in Section 3.1 (which are not essential for any subsequent results), this notation is confined to the Appendixes. At points in the article where some readers may desire more details or clarifications, we explicitly reference the appropriate sections of the article by Mark (2003), where the derivation and discussion of these and other results are presented in a more general, more detailed, and more technical context.

2. FORMAL STATEMENT OF INFERENCE, DATA STRUCTURE, AND SAMPLING PROCESS

2.1 Standardized Survival, Standardized Risk Difference, Full Data, and Auxiliary Covariates

Our main goal of inference is to estimate conditional survivals (3), standardized survivals, $S^s(t|v^\dagger)$, and standardized risk differences. In accord with usual epidemiologic parlance, we define $S^s(t|v^\dagger)$ to be the weighted sum over j of the $S(t|z)$, $z \in \mathcal{Z}$,

$$S^s(t|v^\dagger) = \sum_j S(t|v^\dagger, j^*)w(j^*), \qquad 0 \le t \le \tau. \quad (4)$$

Here $v^\dagger$ and j^* represent specific points in the support of V_i and J_i, and the weights, $w(j^*)$, are functions of j^* chosen by the investigator that sum to 1. In the analysis of the Hp data, the only adjusting covariate, J_i, is age, and we define $w(j^*)$ to be the observed marginal distribution of J_i in the cohort. For a given set of weights, the standardized risk differences are a simple contrast,

$$Rd(t) = S^s(t|v0) - S^s(t|v1), \quad (5)$$

where, for instance, we use $v0$ to represent $V_i = 0$. Because we wish primarily to make inference about survivals in groups of individuals, we assume that the support $\mathcal{Z}$ is finite with k^* levels. Although cumulative hazards and survivals can be estimated at any time t (App. A), for simplicity we always express

these quantities in terms of the end of the study, and for the remainder of the article we set $t = \tau$.

Were resource limitation not a factor, then we could measure V_i on everyone and obtain "full data," $H_i = \{W_i, V_i\}$, where $W_i = \{X_i, \Delta_i, A_i\}$. We assume that the covariates in A_i are measured at time 0; A_i generally contains orders of magnitude more measurements than the set of adjusting covariates, J_i. Although while designing a cohort study it is essential to consider which adjusting covariates will be required, typically J_i is not formally specified until the analysis stage. Then J_i is usually chosen to be the subset of covariates in A_i that are known, or suspected, of being associated with T_i and/or V_i and are not "on the causal pathway." We refer to the (possibly empty) set of covariates that are in A_i but not J_i, as *auxiliary covariates*, Λ_i^{aux}. Thus $A_i = \{J_i, \Lambda_i^{\text{aux}}\}$. The term auxiliary indicates that we do not wish to make inference about the cumulative hazards $\Lambda(t; z)$, conditional on Λ_i^{aux}. In a sense made precise in Section 5 and in the simulations, the Λ_i^{aux} can substantially increase the efficiency of estimation when they are correlated with V_i.

2.2 Stage-Two Sampling Restrictions

We define $R_i = 1$ if V_i is known for individual i, and $R_i = 0$ otherwise. For most of the article, we assume that conditional on W_i, selection of individuals for measurement of V_i is independent with known, nonzero probabilities, $\pi_o(W_i)$, that do not depend on V_i, that is,

$$\pi_o(W_i) = \Pr(R_i = 1|W_i, V_i) = \Pr(R_i = 1|W_i). \tag{6}$$

In the usual parlance of missing data, restriction (6) is consistent with V_i being missing at random (MAR) (Rubin 1976). As we frequently do, we drop the argument of a function and use the subscript i to indicate that the argument is a random variable. Thus we write $\pi_{i,o}$, where $\pi_{i,o} \equiv \pi_o(W_i)$. At the end of Section 4 we extend the results to dependent sampling and to missingness that is not entirely under investigator control.

Without loss of generality, we specify the known sampling probabilities by

$$\operatorname{logit}\pi_o(W_i) = \psi_o^T h(W_i). \tag{7}$$

Here ψ_o and $h(W_i)$ are known, conformable, finite-dimensional vectors of parameters and random variables. It is important to note that the function $h(\cdot)$ is not uniquely determined by the $\pi_o(W_i)$: that is, neither the parameterization nor the dimension of (7) is unique. For instance, if A_i contains only information on sex, and stage-two sampling depends only on case status, then two correctly specified models for (7) are

$$\operatorname{logit}\pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) \tag{8}$$

and

$$\operatorname{logit}\pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) + \psi_{o3} I(\mathit{male}) + \psi_{o4} I(\mathit{female}). \tag{9}$$

Here $\psi_{o1} = \operatorname{logit}\Pr(R_i = 1|\Delta_i = 1)$, $\psi_{o2} = \operatorname{logit}\Pr(R_i = 1|\Delta_i = 0)$, and $\psi_{o3} = \psi_{o4} = 0$.

The usefulness of models such as (9) will become evident when we discuss $\widehat{\pi}$-estimation in Section 4.

We define W_i^R to be the smallest set of linearly independent vectors such that (7) is true, where size refers to the dimension of the column space spanned by $h(W_i)$ [$\operatorname{span} h(W_i)$]. In our earlier example, the dimension of W_i^R is two. Letting $W_i^l \equiv h(W_i)$ for some $h(\cdot)$, correctly specified models are those such that

$$\operatorname{span}(W_i^l) \supseteq \operatorname{span}(W_i^R). \tag{10}$$

We consider models with equivalent spans to be identical, and restrict ourselves to covariate spaces where the W_i^l are linearly independent. We denote the scores from any logistic model with covariates W_i^l as S_i^l,

$$S_i^l = (R_i - \pi_{i,o}) W_i^l. \tag{11}$$

3. ESTIMATORS AND INFLUENCE FUNCTIONS

3.1 Full-Data Estimators and Full-Data Influence Functions

Although our inferential focus is survival, cumulative hazards are the compensators of the counting process, and thus the "natural" scale for estimation. (For this and other details on counting process martingales, see Andersen, Borgan, Gill, and Keiding 1993.)

In full-data studies, the Nelson–Aalen estimator, $\widehat{\Lambda}(\tau; z)$, is the efficient nonparametric estimator of (2) (Andersen et al. 1993). The maximum partial likelihood estimator, $\widehat{\beta}$, and the Breslow estimator, $\widehat{\Lambda}_o(\tau, \widehat{\beta})$, are the semiparametric efficient estimators of β_o and the baseline cumulative hazard (12) (Andersen et al. 1993),

$$\Lambda_o(\tau) = \int_0^\tau \lambda_o(u)\, du. \tag{12}$$

For the semiparametric model, the cumulative hazard at any covariate level z is $\Lambda(\tau; \beta_o z) \equiv \Lambda_o(\tau) \exp(\beta_o^T z)$. It is estimated by replacing the unknown parameters, β_o and $\Lambda_o(\tau)$, with their estimates. To indicate the $k^* \times 1$ vector of cumulative hazards, we drop z from the arguments and write $\Lambda(\tau)$, or $\Lambda(\tau; \beta_o)$. Because we are assuming that the time of interest is τ, we frequently drop the time argument.

Let $\widehat{\alpha}^1$ denote the Nelson–Aalen estimator, and let $\widehat{\alpha}^2$ denote the partial likelihood estimator of $\widehat{\beta}$. Then $\widehat{\alpha}^b$, $b \in \{1, 2\}$, are both solutions to estimating equations of the form

$$\sum_{i=1}^n U_i^b\big(H_i, \mathcal{R}(X_i); \alpha^b\big) = 0. \tag{13}$$

Each term in (13) depends not only on the subjects data, H_i, but also on $\mathcal{R}(X_i)$. $\mathcal{R}(X_i)$ represents the set of individuals at risk at time X_i: $\mathcal{R}(X_i) = \{i : X_j \geq X_i\}$. For instance, using standard counting process notation (see Sec. A.1), when $b = 2$, $\alpha^b = \beta$; $U_i^2(H_i, \mathcal{R}(X_i); \alpha^2) = \int_0^\tau \{Z_i - S^1(u, \beta) S^0(u, \beta)^{-1}\}\, dN_i(u)$, and the maximum partial likelihood estimator, $\widehat{\beta}$, is the β that solves $\sum_{i=1}^n \int_0^\tau \{Z_i - S^1(u, \beta) S^0(u, \beta)^{-1}\}\, dN_i(u) = 0$.

For the Breslow estimator, we first estimate $\widehat{\beta}$. Then, with $\widehat{\alpha}^3$ denoting $\widehat{\Lambda}_o(\tau, \widehat{\beta})$, we can write the estimating equations similarly as

$$\sum_{i=1}^n U_i^3\big(H_i, \mathcal{R}(X_i); \widehat{\beta}; \alpha^3\big) = 0. \tag{14}$$

[See sec. 2 of Mark 2003 for the explicit forms for $b = 1$ and $b = 3$ (14).]

Although the $U_i(\cdot)$ are not iid, the estimators are asymptotically equivalent to a sum of mean 0, independent influence functions (Andersen et al. 1993), that is,

$$n^{1/2}(\widehat{\alpha}^b - \alpha_o^b) = n^{-1/2}\sum_{i=1}^{n} D_i^{Fb}(H_i; \alpha_o^b) + o_p(1), \qquad (15)$$

where α_o^b is the underlying parameter being estimated. We refer to these D_i^{Fb}, $b \in \{1, 2, 3\}$, as the *full-data influence functions* of $\widehat{\Lambda}(\tau)$, $\widehat{\beta}$, and $\widehat{\Lambda}_o(\tau, \widehat{\beta})$. The explicit definitions of the D_i^{Fb} are given in Section A.2. For instance, for $b = 2$, Section A.2 gives $D_i^{F2} = i^{-1}\int_0^\tau \{Z_i - e(u, \beta_o)\}\, dM_i(u)$, and (15) becomes

$$n^{1/2}(\widehat{\beta} - \beta_o) = n^{-1/2}\sum_{i=1}^{n} i^{-1}\int_0^\tau \{Z_i - e(u, \beta_o)\}\, dM_i(u) + o_p(1).$$

3.2 Estimators and Influence Functions for Two-Stage Designs

For two-stage designs, direct application of the theorems in RRZ establish that the solutions to estimating equations

$$\sum_{i=1}^{n} \pi_{i,o}^{-1} R_i U_i^b\big(H_i, \mathcal{R}(X_i); \alpha^b\big) - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_b(W_i) = 0, \quad b \in \{1, 2\}, \qquad (16)$$

produce consistent, asymptotically normal nonparametric, and semiparametric estimators of the cumulative hazards (2) and β_o, respectively. Here the g_b's are any conformable vectors of nonstochastic functions of W_i specified by the investigator. We denote estimators based on the functions g_b as $\tilde{\Lambda}(g_1)$; $\tilde{\beta}(g_2)$.

Similarly, Mark (2003, sec. 4) showed that solutions to

$$\sum_{i=1}^{n} \pi_{i,o}^{-1} R_i U_i^3\big(H_i, \mathcal{R}(X_i); \tilde{\beta}(g_2); \alpha^3\big) - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_3^*(W_i) = 0 \qquad (17)$$

define a class of consistent, asymptotically normal two-stage semiparametric estimators of (12). We denote those estimators by $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$ or $\tilde{\Lambda}_o(\tau, \tilde{\beta}, g_3)$. Here g_3^* is any scalar function of W_i, and $g_3(W_i)$ is the function of g_2 and g_3^* defined in (A.4.1). The explicit estimating equations for (16) and (17) are given in (A.3.1)–(A.3.3).

We write the influence functions that correspond to these classes of two-stage estimators as

$$D_i^b(g_b) = \pi_{i,o}^{-1} R_i D_i^{Fb} - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_b(W_i), \quad b \in \{1, 2, 3\}. \qquad (18)$$

For $b \in \{1, 2\}$, (18) follows directly from RRZ. For $b = 3$, (18) is obtained by a Taylor series expansion around β_o (A.4). Using notation analogous to the full-data case (15), we express the two-stage estimators as a sum of their influence functions,

$$n^{1/2}\big(\tilde{\alpha}^b(g_b) - \alpha_o^b\big) = n^{-1/2}\sum_{i=1}^{n} D_i^b(g_b) + o_p(1). \qquad (19)$$

From (19), it is clear that the asymptotic variances of the $\tilde{\alpha}^b(g_b)$ are $E[D_i^b(g_b) D_i^b(g_b)^T]$.

Let $\tilde{\Lambda}(\tau, z, \cdot)$ be any nonparametric [e.g., $\tilde{\Lambda}(g_1; z)$] or semiparametric [e.g., $\tilde{\Lambda}_o(\tau, \tilde{\beta}, g_3) \times \exp(\tilde{\beta}^T(g_2)z)$] two-stage estimator of (2). Then survival estimates, $\tilde{S}(\tau|z)$, are formed by replacing $\Lambda(\tau; z)$ in (3) with $\tilde{\Lambda}(\tau, z, \cdot)$. Asymptotic distributions are derived by applying the functional delta method exactly as was done by Andersen et al. (1993) for the full data survival estimators. We provide consistent estimators of the variances of two-stage estimators of (3)–(5) in Appendix A and in Mark (2003, app. A.4).

4. THE SIMPLE TRUE-π AND $\widehat{\pi}$-ESTIMATORS

We define *simple true-π* (STP) estimators to be estimators where $g_b \equiv 0$ [e.g., $g_i(W_i) = 0$ for all i]; that is, they are the usual inverse–probability-weighted Horvitz–Thompson estimators. However, rather than using the notation in (18) and denoting the influence function of these STP estimators as $D_i^b(g_b \equiv 0)$, we write $D_i^b(\pi_o)$, which, by (18), is

$$D_i^b(\pi_o) = \pi_{i,o}^{-1} R_i D_i^{Fb}. \qquad (20)$$

We define *$\widehat{\pi}$-estimating procedures* to be procedures in which we continue to set $g_b \equiv 0$, but replace the known $\pi_{i,o}$ in estimating equations (16) and (17) with an estimate, $\widehat{\pi}(W_i^l)$, of $\pi_{i,o}$. The predicted sampling probabilities, $\widehat{\pi}(W_i^l)$, are obtained by replacing ψ_o with its maximum likelihood estimate, $\widehat{\psi}$, in a correctly specified model (7) with covariates $h(W_i) = W_i^l$. We refer to estimators from such procedures as *$\widehat{\pi}$-estimators*. RRZ (prop. 6.2) showed that $\widehat{\pi}$-estimators are consistent and asymptotically normal and have an influence function, $D_i^b(\widehat{\pi}(W^l))$, that is the residual of a population least squares regression of (20) on the scores from the prediction model (11). That is,

$$D_i^b(\widehat{\pi}(W^l)) = D_i^b(\pi_o) - P^{bl} S_i^l. \qquad (21)$$

Here $P^{bl} S_i^l$ is the projection operator,

$$P^{bl} = E[D_i^b(\pi_o) S_i^{l^T}] E[S_i^l S_i^{l^T}]^{-1}. \qquad (22)$$

Because $D_i^b(\widehat{\pi}(W^l))$ is a residual, the variance of the $\widehat{\pi}(W^l)$-estimator is less than or equal to the variance of the STP estimator for all W_i^l. In addition, because residuals are nonincreasing in the dimension of the column space, if $\text{span}(W_i^m) \supset \text{span}(W_i^l)$, then the variance of the $\widehat{\pi}(W^m)$-estimator is less than or equal to the variance of the $\widehat{\pi}(W^l)$-estimator. [For a more in-depth discussion of the properties of $\widehat{\pi}$-estimating procedures, and proof that the $\widehat{\pi}$-estimating procedures and the solutions to estimating equations (16) and (17) generate the identical class of estimators, see Mark 2003, secs. 5 and 6 and app. C.]

$\widehat{\pi}$-estimation is the "natural" estimating procedure when we relax the requirements that sampling be independent with known probabilities. In general, the dependent sampling that we consider is characterized as follows: Partition the observed W_i into a finite number of strata, and select a fixed number of cases and controls from each stratum. If we let W_i^f be the saturated column space of indicator variables generated by that partition, then we can use any $\widehat{\pi}$-estimator with $\text{span}(W_i^l) \supseteq \text{span}\, W_i^f$ (RRZ, lemma 6.2). Such dependent sampling is common; for example, in the Hp study, we sampled a fixed number of cases and controls. NCC risk set sampling is by definition dependent. We review the definition of NCC sampling and provide appropriate $\widehat{\pi}$-estimators in Appendix B.

So far, we have assumed that $\pi_{i,o}$, or equivalently, the ψ_o in logistic models (7), are known. If rather than knowing ψ_o, we only know there is a ψ^* such that $\text{logit}\,\pi_{i,o} = \psi^* W_i^l$, then the estimator $\widehat{\pi}(W^l)$ also has influence function given by (21) (RRZ, prop. 6.2). For instance, to obtain consistent $\widehat{\pi}$-estimators for our Linxian studies, we had to assume that we could correctly specify a logistic model that accounted for the chance missingness. Given the nature of the events causing the missingness (see Sec. 6), we believed that missingness was related to neither W_i or V_i, and hence any $\widehat{\pi}$-estimator with $\text{span}(W_i^l) \supseteq \text{span}(W_i^R)$ would be consistent.

Computer code for implementing the general class of $\widehat{\pi}$-estimating procedures in R is available from the authors. This program handles a completely general data structure and gives estimates, and the variances, for conditional survivals (3), standardized survivals (4), risk differences (5), and population-attributable risks. (For population-attributable risk estimators and their asymptotic distributions, see Mark 2003, app. A.4.) We have used this program to produce nonparametric survival curves for an article analyzing the association of zinc levels in biopsy tissue with esophageal cancer (Abnet et al. 2005), and to produce semiparametric survival curves and risk estimates for the nutrient analyses described in Section 6.

5. EFFICIENCY, IDENTIFIABILITY, AND LOCAL EFFICIENCY

5.1 Efficiency and the Optimal g_b

Referring to $\pi_{i,o}^{-1} R_i$ as the *weight* and $\pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_b(W_i)$ as the *offset*, it is clear from estimating equations (16) and (17) and influence functions (18) that the class of two-stage estimators that we consider comprises *weighted versions with offset* of the efficient full-data estimators. Because specific estimators differ only with regard to the specification of the g_b, efficiency differences are determined entirely by the choice of the g_b function. We use g_b^{eff} to denote the optimal g_b, that is, the g_b that minimizes $E[D_i^b(g_b) D_i^b(g_b)^T]$. By results of RRZ (1994) and Newey (1990), who showed that all regular nonparametric full-data estimators are asymptotically equivalent, the class of nonparametric estimators, $\tilde{\Lambda}(g_1)$, defined by (16) and (A3.1), contains (in the sense of asymptotic equivalence) all possible nonparametric cumulative hazard estimators for two-stage designs. Hence the estimator $\tilde{\Lambda}(g_1^{\text{eff}})$ achieves the nonparametric efficiency bound (Mark 2003). In contrast, for semiparametric estimators, we have followed a "practical recommendation" of RRZ (p. 850) and restricted consideration to the subclass of all possible two-stage semiparametric estimators that use the "full-data efficient $h(\,)$ function" (Mark 2003). [See RRZ for the general definition of the $h(\cdot)$ functions, and its specific form in two-stage estimators of the β_0 in (1).] Thus we call estimators using the g_b that minimizes the variance of D_i^2 and D_i^3 the *restricted-class–efficient* (RC-efficient) estimators (Mark 2003).

For $b \in \{1, 2\}$, direct application of proposition 2.3 of RRZ establishes that $g_{i,b}^{\text{eff}} = E[D_i^{Fb} | W_i]$. By applying the same result to semiparametric estimators (17) of (12), we find that for any given g_2 function, the variance is minimized by $g_{i,3}^* = E[D_i^{F3} | W_i]$. It is simple then to show that the variance of $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), E[D^{F3}|W])$ is minimized with $g_2 \equiv 0$ (Mark 2003, app. B). By definition of g_3 (A.4.1), $g_{i,3}^{\text{eff}} = E[D_i^{F3} | W_i]$.

Replacing $g_b(W_i)$ with $[D_i^{Fb} | W_i]$ in (18) demonstrates that these efficiency results correspond to intuition. Every subject contributes $E[D_i^{Fb} | W_i]$ to estimation; subjects with measured V_i provide the additional information in their observed "weighted residual," $\pi_{i,o}^{-1}(D_i^{Fb} - E[D_i^{Fb} | W_i])$.

One can see that by choosing $g_b(W_i) = \pi_{i,o} P^{bl} W_i^l$, the influence function given by (18) is identical to the influence function of the $\widehat{\pi}$-estimator, $\widehat{\pi}(W_i^l)$ (21). Conversely, for any $g_b(W_i)$ one can specify a logistic model (7) so that the influence function of the $\widehat{\pi}$-estimator is identical to (18) (RRZ; Mark 2003). In particular, $\widehat{\pi}$-estimators based on predicted probabilities from logistic model (23),

$$\text{logit}\,\pi_o(W_i) = \psi_1^T h(W_i) + \psi_2^T W_i^{\text{eff}}, \qquad W_i^{\text{eff}} = \pi_{i,o}^{-1} g_{i,b}^{\text{eff}}, \tag{23}$$

are efficient or RC-efficient (Mark 2003, app. C).

5.2 Identification of g_b^{eff} and Implications for Study Design and Analysis

The optimal $g_{i,b}$, $E[D_i^{Fb} | W_i]$, are functions of unknown parameters. RRZ's proposition 2.4 established that $g_{i,b}^{\text{eff}}$ can be replaced by a consistent estimate, $\widehat{g}_{i,b}^{\text{eff}}$, without changing the asymptotic distributions of the two-stage estimators. That is, an estimator using $\widehat{g}_{i,b}^{\text{eff}}$ achieves the efficiency (or RC-efficiency) bound. If $g_{i,b}^{\text{eff}}$ can be consistently estimated, then we say that the efficient estimator is identified. If not, then the variance of the efficient influence function represents an unknown lower bound that no estimator is guaranteed to achieve.

If the support of W_i were discrete, $g_{i,b}^{\text{eff}}$ could be consistently estimated by the empirical average of the D_i^{Fb} among individuals with $\text{R}_i = 1$ within each level of W_i. Thus a $\widehat{\pi}$-estimator saturated in the discrete W_i obtains that efficiency bound. In time-to-event data, W_i has the continuous component, X_i. Unless X_i is a deterministic function of (Δ_i, A_i), there is no discrete subset $W_i^l \subset W_i$ such that $E[D_i^{Fb} | W_i^l] = E[D_i^{Fb} | W_i]$ (Mark 2003). In the rest of this section we approach the task of conditioning on X_i and increasing efficiency. We do this by reexpressing $g_{i,b}^{\text{eff}}$ in terms of relative risks, survivals, and covariate distributions. We discuss conditions under which each of these can be consistently estimated, and examine the implications for study design and analysis.

We reexpress $g_{i,b}^{\text{eff}}$ (Mark 2003, sec. 8.1) as

$$g_{i,b}^{\text{eff}} = EE[D_i^{Fb} | W_i, V_i] = \int_{\mathcal{V}} D_i^{Fb}(W_i, v) \Pr(v | W_i)\, dv. \tag{24}$$

In the design stage, a crucial consideration is what, if any, auxiliary variables should be measured. From (24), it is clear that for Λ_i^{aux} to be optimal, it is sufficient that for any larger set, $\Lambda_i^{\text{aux}+} > \Lambda_i^{\text{aux}}$,

$$\Pr(v | X_i, \Delta_i, J_i, \Lambda_i^{\text{aux}}) = \Pr(v | X_i, \Delta_i, J_i, \Lambda_i^{\text{aux}+}); \tag{25}$$

that is, we should collect all auxiliary information that provides additional knowledge about the distribution of the incompletely measured covariates V_i at any time on study. Letting v^1 be some reference level of interest in $\mathcal{V}$, we can parameterize the $\Pr(v | W_i)$ in (24) in terms of the *exposure odds*,

$$K_{i,v} = \frac{\Pr(V_i = v | W_i)}{\Pr(V_i = v^1 | W_i)}. \tag{26}$$

Using Bayes's rule, $\Pr(v|W_i) = \Pr(X_i, \Delta_i|v, A_i)\Pr(v|A_i)/\Pr(X_i, \Delta_i|A_i)$, and a noninformative censoring assumption, $\Pr(C_i \geq s|T_i \geq s, V_i, A_i) = \Pr(C_i \geq s|T_i \geq s, A_i)$, (26) becomes (Mark 2003, app. D)

$$K_{i,v} = rr(X_i|v, A_i)^{\Delta_i} \frac{S(X_i|v, A_i)}{S(X_i|v^1, A_i)} \frac{\Pr(v|A_i)}{\Pr(v^1|A_i)}. \quad (27)$$

Here $rr(X_i|v, A_i)$ and $S(X_i|v, A_i)$ are the relative risks and survival times conditional on the event (v, A_i). Then (25) is true if

$$S(u|V_i, J_i, \Lambda_i^{\text{aux}}) = S(u|V_i, J_i, \Lambda_i^{\text{aux}+}), \qquad 0 \leq u \leq \tau, \quad (28)$$

and

$$P(V_i|J_i, \Lambda_i^{\text{aux}}) = P(V_i|J_i, \Lambda_i^{\text{aux}+}). \quad (29)$$

Epidemiologists refer to (28) as Λ_i^{aux} containing all *independent predictors of outcome*, and (29) as Λ_i^{aux} containing all *independent predictors of exposure*.

The requirements for efficient analysis are conceptually and mathematically equivalent to those in the design stage. That is, to estimate $g_{i,b}^{\text{eff}}$, we need only include in the conditioning event the subset of Λ_i^{aux} that contains the independent predictors of outcome and exposure.

5.3 Efficient and Locally Efficient Estimators

Although for any given Λ_i^{aux} it is impossible to know with certainty whether (28) or (29) is true, these are the exact considerations required to control confounding. As described in Section 2, J_i is frequently defined in the analysis stage to be the subset of A_i such that (28) and (29) are "approximately" true when Λ_i^{aux} is removed from the conditioning events on the left side. If successful in selecting all of the disease risk factors in J_i then

$$S(u|V_i, J_i) = S(u|V_i, J_i, \Lambda_i^{\text{aux}}), \qquad 0 \leq u \leq \tau, \quad (30)$$

and (27) becomes

$$K_{i,v} = rr(X_i|v, J_i)^{\Delta_i} \frac{S(X_i|v, J_i)}{S(X_i|v^1, J_i)} \frac{\Pr(v|A_i)}{\Pr(v^1|A_i)}. \quad (31)$$

The identifiability and efficiency results that we give in this section assume that (30) is true.

From (31), it is clear that if we can consistently estimate each term in $K_{i,v}$, we can estimate $\widehat{g}_{i,b}^{\text{eff}}$. For both the nonparametric and semiparametric models, the second and third terms can be estimated by $\widetilde{S}(X_i|v, J_i)$ and $\widehat{P}(v|A_i)$, the empirical average of V_i within levels of A_i. (Here we assume that A_i has finite support; see Mark 2003, app. D, for the case where the support of A_i is not finite.) For the semiparametric model, $rr(u|Z_i)$ can be estimated by $\tilde{rr}(u|Z_i) = \exp(\tilde{\beta}^T Z_i)$. The $\widetilde{S}(X_i|v, J_i)$ and $\tilde{\beta}$ can come from estimates based on any g_2, g_3^* functions. Hence the semiparametric RC-efficient estimators of β_o and $\Lambda_o(\tau)\exp(\beta_o^T Z_i)$ are identified. In contrast, the nonparametric model provides no obvious estimator of $rr(u|Z_i)$. If k^* were small and the number of cases large, then one could theoretically use kernel smooths to estimate hazards, and hence relative risks (rr's). We do not explore this possibility further. Instead, in Section 7.3 we propose several *locally efficient* (LE) estimators. LE estimators approximate g_b^{eff} by making assumptions about $rr(u|Z_i)$. We denote the resultant approximations by $\widehat{g}_{i,b}^{\text{LE}}$. If the assumptions about the rr's are correct, then $\widehat{g}_{i,b}^{\text{LE}}$ is a consistent estimate of $g_{i,b}^{\text{eff}}$, and the LE estimators are efficient. Regardless of the truth of the assumptions, the proposed LE estimators are consistent.

6. TWO–STAGE STUDIES CONDUCTED ON THE LINXIAN COHORT: GOALS, CONSTRAINTS, AND DATA ANALYSIS

6.1 Two-Stage Studies With Cases Missing by Chance

Since 1986, we have been studying a cohort of approximately 30,000 individuals from Linxian, China, a region with epidemic rates of GCC cancer (Blot and Li 1985; Blot et al. 1993). The cohort was assembled to investigate the hypothesis that one or more of the widely prevalent nutrient deficiencies contributed to this high GCC incidence. After following the cohort for 5.25 years and recording data on incident GCC and censoring events, we initiated four major studies where the V_i were measurement(s) of a group of related nutrients (Mark et al. 2000, 2001; Abnet et al. 2003; Taylor et al. 2003). We wanted to estimate nutrient–GCC associations with as much precision as possible, so for these studies our design called for sampling all of the 402 incident GCC cases. Despite the fact that virtually 100% of our cohort consented to giving blood at the beginning of the study in 1986, we discovered that accidents in sample processing, storage, shipping, or laboratory evaluation prevented measurements for approximately 10% of the cancers (Mark et al. 2000). Because this missingness arises from events outside of investigators' control, we refer to such cases as being *missing by chance*. Using the standard case-cohort estimators of relative risk, we found that serum levels of selenium and vitamin E were inversely related to cancer incidence (Mark et al. 2001; Taylor et al. 2003). The strongest effect was for selenium, where individuals in the highest quartile of selenium had approximately half the cancer risk of those in the lowest (Mark et al. 2001). Various strategies for population wide nutrient supplementation to eliminate these deficiencies are currently being considered by our colleagues at the Cancer Institute of the Chinese Academy of Medical Sciences. Decisions of whether and how much to supplement depend on estimates of absolute risks. Using a $\widehat{\pi}$-estimating procedure, we estimated that the correction of both selenium and vitamin E deficiencies could reduce the GCC incidence by approximately 30%. We are currently preparing a manuscript describing these results.

Many of the two-stage studies that we have initiated in the last 4 years have examined the association of GCC with recently characterized DNA polymorphisms (Stolzenberg-Solomon et al. 2003; Savage et al. 2004a,b; Roth et al. 2004; Mahabir et al., in press). Samples suitable for DNA measurements were not collected until 1991, and then only on a subgroup of the remaining cohort. Overall, we measured polymorphisms in approximately 20% of the cases from 1991 to 1996. Thus, in these studies, 80% of the cases were missing by chance.

6.2 "Exploratory" Two-Stage Studies Where Cases Are Missing by Design

By the time we designed the serologic studies of nutrients, numerous other exposures that could be measured in serum had

become of interest. Because our total serum quantity was quite limited and the list of exposures of interest was large, we initiated "exploratory" two-stage studies in which we deliberately sampled only a fraction of cases (Abnet et al. 2001; Limburg et al. 2001). Our goal was to sample only the number of cases and controls required to produce sufficiently precise estimates of exposure prevalence, assay reliability, and risk magnitude to determine whether to commit additional resources (Mark and Katki 2001). In one "preliminary study" where the exposure was the fungal-produced toxin fumonosin, we found that the newly developed measurement procedure was not reliable, and have not initiated a larger study (Abnet et al. 2001). In contrast, due to the results from the "exploratory" study on the association of GCC with serologic evidence of Hp infection (Limburg et al. 2001), we have begun a much larger two-stage study. In studies such as these where V_i is deliberately measured on only a fraction of the cases, we say that the cases are *missing by design*.

6.3 Background Information on the Association of Hp With GCC

Cancers that arise in the proximal 2–3 cm of the stomach are called GCCs. These differ with regard to population rates, and some individual-level risk factors, from stomach cancers that arise outside of the cardia (GNC) (Devesa, Blot, and Fraumeni 1998). In the last decade, epidemiologic cohort studies have found that individuals with Hp infection are at increased risk for GNC; relative risks (*rr*'s) range from 2 to 4 (Helicobacter and Cancer Collaborative Group 2001). The quantity, consistency, and biologic plausibility of the evidence is such that Hp is categorized as a class 1 human carcinogen (International Agency for Research on Cancer 1994).

Prior to our study, only a few small studies, with case sizes ranging from 4 to 12, examined the Hp–GCC association. All were from first-world Western nations. The consensus was that Hp was "protective" for GCC, with $rr \approx .5$ (Helicobacter and Cancer Collaborative Group 2001; Dawsey, Mark, Taylor, and Limburg 2002). Various mechanistic hypotheses have been advanced to account for the opposite association of Hp on GNC and GCC (Blaser 1999).

6.4 Design and Analysis of the Hp–GCC Study Using a Cohort From Linxian, China

Based on dissimilarities between the populations, and on differences in the prevalence of esophageal adenocarcinomas, a type of cancer which can be difficult to distinguish from GCC (Limburg et al. 2001; Dawsey et al. 2002), we hypothesized that the Hp–GCC association in Linxian might differ from that found in Western populations. In accord with the goals for "exploratory" studies given earlier, we sampled approximately 25% of the GCC cases (100 cases) and 7% of controls (200 controls) that occurred in the cohort by $\tau = 5.25$ years (Limburg et al. 2001). We measured serum antibodies and found an Hp prevalence (Hp^+, $V_i = 1$) of approximately 65% and an *rr* of approximately 2 for Hp^+ individuals. The only other major independent risk factor for GCC in this population was age; age greater than the cohort median age, ($J_i = 1$) increased GCC risk by a factor of 3.5.

Table 1 contains estimates of covariate-specific survivals (3), age-standardized survivals (4), and risk differences (6) based on the CPH model (1) with V_i and J_i indicator variables. Because a fixed number of cases ($n = 200$) and controls ($n = 100$) were sampled, we used a $\widehat{\pi}$-estimator to estimate both β_o and $\Lambda_o(\tau)$. In particular, we used logistic model (9), the model saturated in (Δ_i, J_i). Throughout this article, we denote this estimator by $\widehat{\pi}(\Delta, J)$. At each age level, the Hp^+ group had lower survival than the Hp^- group. Within levels of Hp exposure, survival was higher in the younger group (J0). We estimated the age-standardized risk difference to be 1.08%, with a 95% confidence interval whose lower limit just excludes 0.

We contributed the data from our study to a pooled study examining Hp and gastric cancer risks. The overall conclusion of that analysis was that there was no evidence of an Hp–GCC association (Helicobacter and Cancer Collaborative Group 2001). We did not share that interpretation. Rather, we argued that tests for heterogeneity of risk estimates by geographic region were highly significant (Dawsey et al. 2002), and that pooling the risk estimates from Western populations and Chinese populations was not appropriate. We have currently initiated a larger study sampling from the approximately 1,000 GCCs that accrued through 2001 ($\tau = 15$ years). This is also a study in which cases are missing by design; however, here the motivation for the designed missingness is opposite to that described earlier. Based on the Hp prevalence and risk estimates from the "exploratory" study, we determined that measurements on all 1,000 GCCs were not needed to achieve the precision required to eliminate type 1 error as a viable explanation for our earlier findings. The simulations in Section 7 are based on the structure of this new study. We used similar simulations to help arrive at the sampling fractions used in the actual study.

7. SIMULATIONS

7.1 Simulation Parameters and Definition of Relative Efficiency

For all simulations, the marginal covariate probabilities were $\Pr(J1) = .5$ and $\Pr(V1) = .65$. T_i was specified by CPH model (1), $\lambda_o(u)$ was exponential, and censoring was independent. The magnitudes for the baseline hazard and competing

Table 1. Effect of H. pylori *Infection on Age-Specific Survival and Age-Standardized Survival, at 5.25 Years in the Linxian Cohort*

	Hp^- (V0)	Hp^+ (V1)
Young (J0)	99.2% (98.9, 99.5)	98.8% (98.4, 99.0)
Old (J1)	97.3% (96.1, 98.1)	95.5% (94.4, 96.3)
Age-standardized survival	98.3% (97.7, 98.9)	97.2% (96.7, 97.8)
Age-standardized risk difference	1.08% (.02, 2.15)	

NOTE: The estimates are based on CPH model (1) with relative risks $\exp(\beta_{o1} V_i + \beta_{o2} J_i)$.

Table 2. Relative Efficiencies of the $\hat{\pi}(\Delta)$ and $\hat{\pi}(\Delta, J)$ Semiparametric Estimators of $S(\tau|v)$ When J Is an Auxiliary Covariate

	Relative efficiency $rr_V = 2.0$				Relative efficiency $rr_V = .5$			
	$S(\tau\|v0) = 90\%$		$S(\tau\|v1) = 81\%$		$S(\tau\|v0) = 90\%$		$S(\tau\|v1) = 95\%$	
$P(V1\|J1)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$
.65	82	82	46	46	68	68	62	63
.75	81	79	47	46	67	64	62	61
.85	81	73	47	43	66	58	64	57
.95	80	62	47	40	64	48	65	49

NOTE: Relative efficiency equals 100 times the ratio of the variance of an estimator to the variance of the STP estimator. Marginal covariate probabilities are $P(V1) = .65$ and $P(J1) = .5$.

risks were chosen to produce approximately 1,000 expected cases in a cohort of size $n = 6{,}600$ by time τ. Unless noted otherwise, $\exp(\beta_{o1}) = 2$ ($rr_v = 2$). The V–J association was altered by changing the conditional probabilities, $P(V1|J1)$. Stage 2 sampling was binomial and depended only on case status (8). Control sampling was 15%. For the simulations in Tables 2 and 3, 25% of the cases were sampled, resulting in a control: case ratio of approximately 3:1. In Figures 1 and 2, case sampling percentages are indicated along the x-axis.

Each of the results represents the average of 2,000 realizations. All estimators of survivals and β_o were unbiased (the mean of the estimators was always within .1% of the truth). The coverage for 95% confidence intervals ranged from 93.4% to 95.8%. Consequently, rather than present the estimator-specific averages in the tables, we report only relative efficiencies (REs), which we define as the ratio ($\times 100$) of the variance of a given estimator to the variance of the STP estimator. The smaller the RE, the greater the efficiency. Because our focus is on survival estimation, we do not report the REs of the estimators of β_o.

7.2 STP, RC-Efficient, and $\widehat{\pi}$ Semiparametric Estimators of Survival

The data in Table 2 were generated from CPH model (1) with $\beta_{o2} = 0$; $S(\tau|v)$ was estimated by fitting the one-covariate CPH model, $\lambda_o(u)\exp(\beta_{o1}V_i)$. For simulations on the left side of Table 2, $rr_v = 2$, $S(\tau|v0) = 90\%$, and $S(\tau|v1) = 81\%$. For simulations on the right, $rr_v = .5$, $S(\tau|v0) = 90\%$, and $S(\tau|v1) = 95\%$. In these simulations J_i is an auxiliary variable rather than a risk factor. For example, J_i might be a surrogate for V_i, such as evidence of gastric inflammation found in a biopsy specimen obtained at the beginning of the study. We compare $\widehat{\pi}(\Delta)$ estimators based on logistic model (8) with the $\widehat{\pi}(\Delta, J)$ estimator based on (9) at five different levels of V–J association. We first focus on the $rr_v = 2$ simulations.

When $\Pr(V1|J1) = .65$, V_i and J_i are independent. Hence the $\widehat{\pi}(\Delta)$ and $\widehat{\pi}(\Delta, J)$ estimators are equally efficient, and are considerably more efficient than the STP estimator. Because the $\widehat{\pi}(\Delta)$ estimator makes no use of the auxiliary variable, its RE is unchanged as $\Pr(V1|J1)$ increases. In contrast, the efficiency of the $\widehat{\pi}(\Delta, J)$ estimator increases (i.e., RE decreases).

Differences in the efficiencies between two-stage estimators are determined largely by the extent to which information from cases with $R_i = 0$ is used. In the $rr_v = 2$ simulations, the magnitude of the efficiency gains for both $\widehat{\pi}$-estimators is greater in the $v1$ strata than in the $v0$ strata. These greater gains reflect the fact that there are more cases, and hence more cases

Table 3. Relative Efficiency of RC-Efficient, Locally Efficient, and $\hat{\pi}$-Estimators of Standardized Survivals for Semiparametric and Nonparametric Models

		Relative efficiency	
	Estimator	$S^s(\tau\|v0) = 90.4$	$S^s(\tau\|v1) = 82.0$
Semiparametric	$\tilde{S}^{\text{eff}}$	86	41
	$\hat{\pi}(\Delta, J)$	87	45
Nonparametric	SLE correct	90	42
	ILE correct	90	42
	SLE prior	108	47
	ILE prior	90	42
	SLE null	91	43
	ILE null	90	42
	$\hat{\pi}(\Delta, J)$	90	47

NOTE: Relative efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator. Marginal covariate probabilities are $P(V = 1) = .65$ and $P(J = 1) = .5$, with $P(V = 1|J = 1) = .85$ and $P(V = 1|J = 0) = .45$.

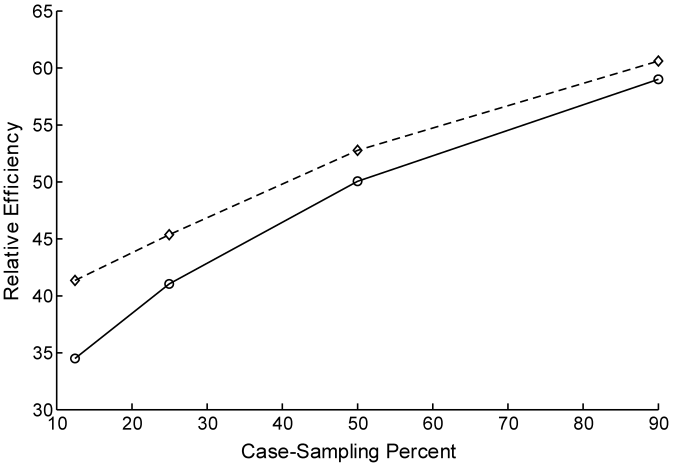

Figure 1. Relative Efficiency of Semiparametric Estimators of $S^s(\tau|v1)$ as Case-Sampling Percentage Varies [---- $\hat{\pi}(\Delta, J)$ estimator; —— $\tilde{S}^{\text{eff}}$ estimator]. The simulation data were generated using the same CPH model as in Table 3. Case-sampling percent is the binomial probability ($\times 100$) of V measurement for cases, and is indicated on the x-axis. For all simulations, control sampling is 15%. Relative efficiency is defined as 100 times the ratio of the variance of the $\hat{\pi}(\Delta, J)$ and $\tilde{S}^{\text{eff}}$ (RC-efficient) estimators, to the variance of the STP estimator. Both estimators are substantially more efficient than the STP estimator. The magnitude of the efficiency gains are inversely related to case-sampling percent. Efficiency differences between the $\tilde{S}^{\text{eff}}$ and $\hat{\pi}(\Delta, J)$ estimators show a similar dependency on case-sampling percent.

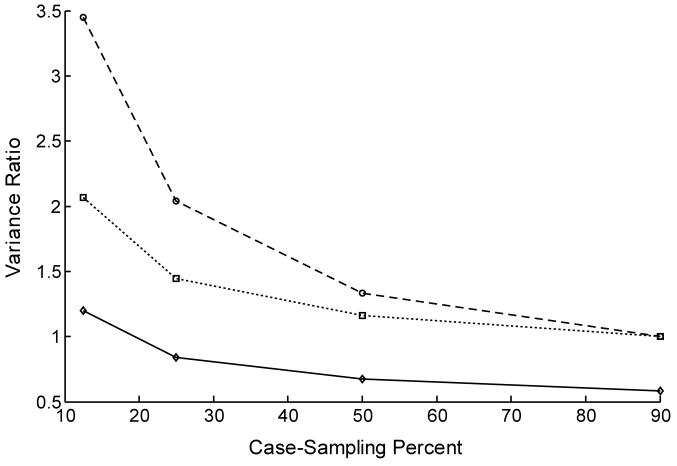

Figure 2. Comparing the Effect of Case-Sampling Percent on the Variance of STP and $\tilde{S}^{eff}$ Estimators of $S^s(\tau|v1)$. The simulation data were generated using the same CPH model as in Table 3. Sampling percent is the binomial sampling probability (×100) of V measurement. For all simulations, control sampling is 15%. Case-sampling percent is indicated on the x-axis. The solid line is the ratio of the variance of the $\tilde{S}^{eff}$ estimator at each of four case sampling percents, to the variance of the STP estimator at 90% case sampling. Except at the lowest case-sampling percent, 12.5%, the ratio is <1. The dotted line compares the variance of the $\tilde{S}^{eff}$ estimator at each case sampling percent, to the variance of the $\tilde{S}^{eff}$ estimator at 90% case sampling. The dashed line plots the corresponding ratios for the STP estimator. The variances of both estimators increase as case sampling fractions decrease. The rate of increase is greater for the STP estimator.

with missing measurement, in the $v1$ stratum. When $rr_v = 2$, 78% of the missing cases occur in the $v1$ group. In contrast, when $rr_v = .5$, 49% of the missing cases occur in the $v1$ stratum. Here both strata have nearly identical REs. Contrasting the two sets of simulations in regard to estimation of $S(\tau|v1)$, we find lower REs when $rr_v = 2$. The expected number of missing cases is 775 for the $rr_v = 2$ simulation and 209 for the $rr_v = .5$.

The comparable nonparametric $\widehat{\pi}$-estimators for the Table 2 simulations produced the same patterns and are not shown. Variances of the nonparametric estimators were 5–10% larger than their semiparametric counterparts.

The first two rows of Table 3 contain results for the RC-efficient estimator, which we denote by $\tilde{S}^{\text{eff}}$, and the $\widehat{\pi}(\Delta, J)$ estimator of standardized survivals for the semiparametric model (1) with $\exp(\beta_{o1}) = 2$ and $\exp(\beta_{o2}) = 3$. As expected, the REs of both estimators are less than one. The $\tilde{S}^{\text{eff}}$ estimator, which conditions on all of W_i when estimating $E[D_i^{Fb}|W_i]$, is more efficient than the $\widehat{\pi}(\Delta, J)$ estimator which conditions only on (Δ_i, J_i). Figure 1 presents the relative efficiencies of the $\tilde{S}^{\text{eff}}$ and $\widehat{\pi}(\Delta, J)$ estimators of $S^s(\tau|v1)$ at four different case sampling probabilities: 12.5%, 25%, 50%, and 90%. The positive slopes of both lines indicate that the efficiency gains decrease as the sampling fraction increases. At 12.5% case sampling, the RE of the $\tilde{S}^{\text{eff}}$ estimator is 35%; at 90% case sampling, the RE is 59%. Similarly, differences in efficiency between the $\tilde{S}^{\text{eff}}$ (solid line) and $\widehat{\pi}(\Delta, J)$ (dotted line) estimators diminish with increasing case sampling percentage. Both of these findings are in accordance with the previous observation that efficiency gains depend on the number of missing cases. The same explanation accounts for the greater efficiency gains for estimators of $S^s(\tau|v1)$ compared with estimators of $S^s(\tau|v0)$ in Table 3; 85% of the expected 1,023 cases have $V_i = 1$.

In Figure 2 the solid line is a plot of the ratio of the variance of the $\tilde{S}^{\text{eff}}$ estimator of $S^s(\tau|v1)$ at each of the four case sampling probabilities to the variance of the STP estimator at 90% sampling. The RC-efficient estimator has a lower variance when 25% of the cases are sampled than the STP estimator has when 90% of the cases are sampled. The dotted line compares the variance of $\tilde{S}^{\text{eff}}$ at each case sampling percent to the variance of $\tilde{S}^{\text{eff}}$ at 90%. The dashed line plots the corresponding ratios for the STP estimator. Although clearly the variances increase as case sampling decreases for both estimators, the rate of increase is greater for the STP estimator. An STP estimator loses all the information from each missing case; the $\tilde{S}^{\text{eff}}$ estimator retains the information contained in $E[D_i^{F3}|W_i]$.

7.3 STP, $\widehat{\pi}(\Delta, J)$, and Locally Efficient Nonparametric Estimators of Survival

The last seven rows of Table 3 contain efficiency results for seven nonparametric estimators of survival. There are three *simple local efficient* estimators (SLEs), three *insured local efficient* estimators (ILEs), and the $\widehat{\pi}(\Delta, J)$ estimator. Each of the corresponding SLEs and ILEs uses identical estimates, $\widehat{g}_1^{\text{LE}}$, of g_1. The distinction is that the SLE estimates are produced by setting $g_{i,1} = \widehat{g}_{i,1}^{\text{LE}}$ in (16) [or (A.3.1)], whereas the ILE estimators are $\widehat{\pi}$-estimators based on prediction model (23) with $g_{i,1}^{\text{eff}}$ replaced by $\widehat{g}_{i,1}^{\text{LE}}$. By construction, ILE estimators must be at least as efficient as $\widehat{\pi}(\Delta, J)$ estimators, even when $\widehat{g}_{i,1}^{\text{LE}}$ is based on a misspecified $rr(X_i|v)$ (Mark 2003). SLEs do not share this property. For example, for the SLE-correct and ILE-correct estimators, the exposure odds (26) are estimated by replacing the relative risks in (31) by estimates from a correctly specified model of the relative risks. Specifically, we assumed exponential hazards within each V_i level, estimated the hazards by dividing the number of observed cases by total person-time, and estimated $rr(X_i|v1)$ as a ratio of the hazards. Both the SLE- and ILE-correct estimators attain the nonparametric efficiency bound. In contrast, for the SLE- and ILE-*prior* and *null* survival estimators the estimates of $rr(X_i|v)$ in (31) are based on misspecified models for the relative risks. The SLE- and ILE-prior estimators set $rr(X_i|v1) = .5$. This is the pooled estimate of relative risk from the prior studies (Sec. 6.3). The SLE- and ILE-null estimators set $rr(X_i|v1) = 1$. These are the efficient estimators under the null hypothesis. Table 3 shows that for estimators of $S^s(\tau|v0)$, not only is the SLE-prior estimator less efficient than the $\widehat{\pi}(\Delta, J)$ estimator, it is also 8% less efficient than the STP estimator. In contrast, the ILE-prior and ILE-null estimators are as efficient (to two significant digits) as the ILE-correct estimator.

8. DISCUSSION

Two-stage studies are commonly used in epidemiology as a resource-effective means of estimating the association of disease with exposures whose measurements consume a substrate that is limited in quantity. When estimating survival, the procedures proposed by the case-cohort and nested case-control designs are biased if cases are missing exposure measurements. In this article, referring to our Linxian studies as examples, we

have described how case missingness arises regardless of investigator intent and explained why designs that deliberately sample a fraction of cases are frequently desirable. Applying results of RRZ, we have derived a class of nonparametric estimators and a class of semiparametric estimators that provide unbiased estimates of cumulative hazards and survivals when cases are missing covariate data (Mark 2003). We used a semiparametric estimator to analyze data from a study in which only 25% of cases were sampled, and found significant differences in age-standardized survivals between subjects with and without serologic evidence of *H. pylori* infection.

Through simulations, we have demonstrated that the variation in efficiency between estimators within a class is of practical consequence. Efficient estimators make better use of the data observed in stage one to provide information on the exposures not observed in stage two. We express the optimal estimators in terms of the familiar quantities of relative risks, survivals, and exposure prevalence, and provide practical strategies for using this formulation to construct estimators with desirable properties. In the design stage, efficiency considerations require collecting information on all covariates suspected of being independent predictors of either exposure or disease. For the analysis stage, we provide a robust procedure ($\widehat{\pi}$-estimation) that incorporates these independent predictors into estimation. R code for implementing these procedures for two-stage studies is available on request.

Given the ease of implementation and the considerable efficiency advantages that even the simplest $\widehat{\pi}$-estimators have compared with the Horvitz–Thompson (STP) estimator, we recommend that the latter never be used for estimating survival from two-stage designs.

APPENDIX A: ESTIMATING EQUATIONS, INFLUENCE FUNCTIONS, AND VARIANCE ESTIMATORS

Here we explicitly define random variables for a univariate counting process such as is appropriate for the semiparametric estimators with CPH model (1). For nonparametric estimators, or semiparametric estimators with a stratified CPH model, the processes should be interpreted in terms of the standard multivariate extension (Andersen et al. 1993; Mark 2003). For instance, in nonparametric estimation $\tilde{\Lambda}(\tau, g_1)$ is the $k^* \times 1$ estimator with row entries $\Lambda(\tau, g_1; z)$ and $N_i(u)$ is $k^* \times 1$, with the kth row defined as $N_{ik}(u) = 1$ iff $I(Z_i = k)$, $T_i \le u$, and $T_i \le C_i$. When we can do so without confusion, and to indicate that any consistent estimator of a parameter will suffice, we drop the argument g_b from two-stage estimators and influence functions; for example, we write $\tilde{\Lambda}(\tau)$ for $\tilde{\Lambda}(\tau, g_1)$. To estimate cumulative hazards and survivals at some time $t \ne \tau$, substitute t for τ in the upper limit of the integrals that define the cumulative hazard estimators.

A.1 Definitions of Counting Process Notation

$$N_i(u) = 1 \quad \text{iff } T_i \le u, \text{ and } T_i \le C_i;$$

$$Y_i(u) = 1, \quad \text{iff } (C_i \wedge T_i) \le u.$$

$$dM_i(u) = dN_i(u) - d\Lambda_i(u); \qquad d\Lambda_i(u) = Y_i(u)\lambda(u|Z_i).$$

$$S^0(u) = \sum_{j=1}^{n} R_i Y_i(u); \qquad S^0(u, \beta) = \sum_{i=1}^{n} Y_i(u) \exp(\widehat{\beta}^T Z_i).$$

$$d\tilde{M}_i(u) = dN_i(u) - Y_i(u)\, d\tilde{\Lambda}(u);$$

$$d\tilde{M}_i(u, \beta) = dN_i(u) - Y_i(u)\, d\tilde{\Lambda}_o(u, \tilde{\beta}) \exp(\tilde{\beta}^T Z_i).$$

$$\tilde{S}^0(u) = \sum_{j=1}^{n} \pi_{i,o}^{-1} R_i Y_i(u);$$

$$\tilde{S}^0(u, \beta) = \sum_{i=1}^{n} \pi_{i,o}^{-1} R_i Y_i(u) \exp(\tilde{\beta}^T Z_i);$$

$$\tilde{S}^1(u, \beta) = \sum_{i=1}^{n} Y_i(u) Z_i \exp(\tilde{\beta}^T Z_i);$$

$$\tilde{E}(u, \beta) = \tilde{S}^1(u, \beta)\tilde{S}^0(u, \beta)^{-1}.$$

$$\tilde{i} = n^{-1} \sum_{i=1}^{n} \pi_{i,o}^{-1} R_i \Delta_i \big(Z_i - \tilde{E}(X_i, \beta)\big)\big(Z_i - \tilde{E}(X_i, \beta)\big)^T;$$

$$n^{-1}\tilde{S}^j(u, \cdot) \xrightarrow{\text{limp}} s^j(u, \cdot) \quad \text{for } j \in \{0, 1\};$$

$$\tilde{E}(u, \beta) \xrightarrow{\text{limp}} e(u, \beta_o) = s^1(u, \beta_o)s^0(u, \beta_o)^{-1};$$

$$\tilde{i} \xrightarrow{\text{limp}} i = E\left[\left(\int_0^\tau \{Z_i - e(u, \beta_o)\}\, dM_i(u)\right) \times \left(\int_0^\tau \{Z_i - e(u, \beta_o)\}\, dM_i(u)\right)^T\right].$$

Here $\xrightarrow{\text{limp}}$ means limit in probability. (For more details, see Andersen et al. 1993; for weighted processes, see Pugh 1993.)

A.2 D_i^{Fb}: The Full-Data Influence Functions (Andersen et al. 1993)

$$D_i^{F1} = \int_0^\tau [s^0(u)]^{-1}\, dM_i(u);$$

$$D_i^{F2} = i^{-1} \int_0^\tau \{Z_i - e(u, \beta_o)\}\, dM_i(u);$$

$$D_i^{F3} = \int_0^\tau [s^0(u, \beta_o)]^{-1}\, dM_i(u) - D_i^{F2^T} \int_0^\tau e(u, \beta_o)\, d\Lambda_o(u, \beta_o).$$

A.3 Two-Stage Estimators of $\Lambda(\tau)$, β_o, $\Lambda_o(\tau)$

$$\tilde{\Lambda}(\tau, g_1) = \sum_{i=1}^{n} \left\{ \pi_{i,o}^{-1} R_i \int_0^\tau \big(\tilde{S}^0(u)\big)^{-1}\, dN_i(u) - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_1(W_i) \right\}. \quad \text{(A.3.1)}$$

$\tilde{\beta}(g_2)$ is the β that solves

$$\sum_{i=1}^{n} \left\{ \int_0^\tau \pi_{i,o}^{-1} R_i \big(Z_i - \tilde{E}(u, \beta)\big)\, dN_i(u) - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_2(W_i) \right\} = 0. \quad \text{(A.3.2)}$$

To estimate $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$, first estimate $\tilde{\beta}(g_2)$ in (A.3.2); then

$$\tilde{\Lambda}_o\big(\tau, \tilde{\beta}(g_2), g_3^*\big) = \sum_{i=1}^{n} \left\{ \int_0^\tau \pi_{i,o}^{-1} R_i \big[\tilde{S}^0\big(u, \tilde{\beta}(g_2)\big)\big]^{-1}\, dN_i(u) - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_3^*(W_i) \right\}. \quad \text{(A.3.3)}$$

A.4 Deriving the Influence Function for $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$

To show that when $b = 3$, (18) is the influence function for $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$, we write

$$\begin{aligned}&\tilde{\Lambda}_o\big(\tau, \tilde{\beta}(g_2), g_3^*\big) - \Lambda_o(\tau, \beta_o)\\ &\quad= \big\{\tilde{\Lambda}_o\big(\tau, \tilde{\beta}(g_2), g_3^*\big) - \tilde{\Lambda}_o(\tau, \beta_o, g_3^*)\big\}\\ &\qquad+ \big\{\tilde{\Lambda}_o(\tau, \beta_o, g_3^*) - \Lambda_o(\tau, \beta_o)\big\}.\end{aligned}$$

Using a Taylor series expansion of $\tilde{\beta}(g_2)$ around β_o as in theorem VII 2.3 Andersen et al. (1993), the first term on the right side is $(\tilde{\beta}(g_2) - \beta_o)^T \int_0^\tau e(u, \beta_o)\lambda_o(u)\,du + o_p(1)$. Multiplying by $n^{1/2}$ and replacing estimators with their influence functions gives

$$\begin{aligned}D_i^3(g_3) &= \pi_{i,o}^{-1} R_i D_i^{F3} - \pi_{i,o}^{-1}(R_i - \pi_{i,o}) g_3(W_i);\\ g_{i,3} &= g_{i,3}^* - g_{i,2}\int_0^\tau e(u, \beta_o)\,d\Lambda_o(u, \beta_o).\end{aligned} \tag{A.4.1}$$

A.5 Estimating $D_i^b(g_b)$ (18) and $D_i^b(\widehat{\pi}(W^l))$ (21)

Estimators $\tilde{D}_i^b(g_b)$ of $D_i^b(g_b)$ are formed by replacing the $s^j(u, \cdot)$, $dM_i(u, \cdot)$, and $e(u, \beta)$ in the D_i^{Fb} with their estimators in Section A.1. The weights $\pi_{i,o}$ can be replaced by any consistent estimate, $\widehat{\pi}$. For $\widehat{\pi}$-estimation, estimators $\tilde{D}_i^1(\widehat{\pi}(W^l))$ and $\tilde{D}_i^2(\widehat{\pi}(W^l))$ are the residuals from an ordinary least squares regression of $\tilde{D}_i^b(\pi_o)$ on the scores $\tilde{S}_i^l$. For $b = 3$, the influence function (21) is correct for $\widehat{\pi}$-estimator where $g_2 = 0$. For $\widehat{\pi}$-estimators with any $\tilde{\beta}(g_2)$ used in (A.3.3), the influence function is

$$\begin{aligned}&D_i^3\big(g_2, \widehat{\pi}(W^l)\big)\\ &\quad\equiv D_i^3(g_2, g_3^* = 0) - E\big[D_i^3(g_2, g_3^* = 0)S_i^{l^T}\big]E[S_i^l S_i^{l^T}]^{-1}S_i^l.\end{aligned} \tag{A.5}$$

Equation (A.5) is derived by sequential application of RRZ's proposition 6.2 (Mark 2003). (A.5) is estimated by least squares regression residuals as described earlier. In the particular instance in which the estimates of β_o come from $\widehat{\pi}$-estimation, $g_{i,2}(\pi) = \pi_{i,o}P^{2l}W_i^l$ (Mark 2003).

A.6 Estimating the Asymptotic Variance of $\tilde{\Lambda}(\tau)$ and $\{\tilde{\beta}^T, \Lambda_o(\tau)\}^T$

Let $\tilde{D}_i^a = \{(\tilde{D}_i^2)^T, \tilde{D}_i^3\}^T$, and let V_1 and V_a be the variances of $\tilde{\Lambda}(\tau)$ and $\{\tilde{\beta}^T, \tilde{\Lambda}_o(\tau, \tilde{\beta})\}^T$. Consistent estimates of the asymptotic variance are $\tilde{V}_1 = n^{-1}\sum \tilde{D}_i^1 \tilde{D}_i^{1T}$ and $\tilde{V}_a = n^{-1}\sum \tilde{D}_i^a \tilde{D}_i^{aT}$.

A.7 Estimating the Asymptotic Variances of $\tilde{S}(\tau|z)$, $\tilde{S}^s(\tau|v)$, and $\tilde{R}d(\tau)$

Let $\tilde{S}(\tau)$ and $\tilde{S}(\tau, \beta)$ be the $k^* \times 1$ vector of nonparametric and semiparametric estimates of $S(\tau)$, with row h entry $\tilde{S}(\tau|h)$ and $\tilde{S}(\tau|h, \beta)$; here h is a point in the support of Z_i. Let V_{s1} and V_{s2} be the corresponding $k^* \times k^*$ variance matrices for $\tilde{S}(\tau)$ and $\tilde{S}(\tau, \beta)$. Define G as the $k^* \times k^*$ diagonal matrix with $\tilde{S}(h)$ in the hth row, hth column. Then $\tilde{V}_{s1} = G\tilde{V}_1 G_1$ is a consistent estimate of V_{s1}. Each h can be represented as a unique $p \times 1$ covariate vector, z_h. Let $L_h = \tilde{S}(\tau|h, \beta)\exp(\tilde{\beta}^T z_h) \times \{1, \tilde{\Lambda}_o(\tau, \beta) \times z_h\}$. Let L be the $k^* \times (p + 1)$ matrix with hth row L_h^T. Then $\tilde{V}_{s2} = L\tilde{V}_a L^T$ is a consistent estimator of V_{s2}.

Let v^* and j^* be the number of levels in the support of V_i and J_i respectively. Let $\tilde{S}(\tau, \cdot)$ be either the nonparametric or semiparametric estimator of $S(\tau)$. Arrange $\tilde{S}(\tau, \cdot)$ in v^* groups of length j^*, in order of increasing index. Let W_j^T be the $1 \times j^*$ matrix of weights w_j, let I_{v^*} the $v^* \times v^*$ identity matrix, and let $C_w = W_j^T \otimes I_{v^*}$, where $\otimes$ denotes the Kronecker product. Then $\tilde{S}^s(\tau|v) = C_w\tilde{S}(\tau, \cdot)$, with variance estimated by, for instance, $C_w\tilde{V}_{s1}C_w^T$. Estimates of standardized risk differences, $\tilde{R}d(\tau)$, are simple contrasts of the $\tilde{S}^s(\tau|v)$. (For estimators of population attributable risk and their distribution, see Mark 2003, app. A.)

APPENDIX B: $\widehat{\pi}$-ESTIMATORS FOR CASE–COHORT AND NESTED CASE–CONTROL DESIGNS

In this section we provide $\widehat{\pi}$-estimators when sampling follows that defined by either the CCH or NCC designs. For simplicity, we assume that sampling does not depend on A_i. Although both designs specify that V_i be observed on all cases, the $\widehat{\pi}$-estimators that we give require no such restriction. We assume only that cases are sampled with some known (dependent or independent probability). (For detailed descriptions of sampling procedures, see, e.g., Self and Prentice 1988; Borgan et al. 1995.)

In the CCH, the "comparison" group is a binomial random sample drawn from all cohort members. Because both the case and the control sampling probabilities are dependent only on Δ_i, any $\widehat{\pi}$-estimators with column space greater than (8) can be used.

NCC designs use dependent, risk set sampling. Let $\{T_{(1)}, \ldots, T_{(d)}\}$ be the set of ordered case failure times. We can estimate the case sampling probability, $\pi_{i,o}(\Delta_1)$, by the proportion of cases sampled. For subjects with $\Delta_i = 0$, we define indicator variables, $R_{ik} = 1$, if the subject is selected at $T_{(k)}$ and $\overline{R}_{ik} = 1$ if $R_{ih} = 1$, for some $h \le k$; $\overline{R}_{i0} \equiv 0$. Let $\pi_{i,k} \equiv \Pr(R_{ik} = 1|X_i, \Delta_i = 0, \overline{R}_{ik-1} = 0)$; then

$$\begin{aligned}\Pr(R_i = 1|\Delta_i = 0, X_i) &\equiv \pi_{i,o}(\Delta_0)\\ &= \sum_{k=1}^{d} \pi_{ik} I\big(X_i \ge T_{(k)}, \overline{R}_{ik-1} = 0\big)\prod_{j=1}^{k-1}(1 - \pi_{ij}),\end{aligned} \tag{B.1}$$

where the product term is defined to be 1 when $k = 1$. To estimate $\pi_{i,o}(\Delta_0)$, we replace the π_{ik} in (B.1) with the proportion of controls with $(X_i \ge T_{(k)}, \overline{R}_{ik-1} = 0)$ who were sampled at $T_{(k)}$. Though (21) remains the correct expression for the influence function, the scores are from a likelihood based on (B.1).

[Received November 2003. Revised June 2005.]

REFERENCES

Abnet, C. C., Borkowf, C. B., Qiao, Y. L., Albert, P. S., Wang, E., Merrill, A. H., Mark, S. D., Dong, Z. W., Taylor, P. R., and Dawsey, S. M. (2001), "Sphingolipids as Biomarkers of Fumonosin Exposure and Risk of Esophageal Squamous Cell Carcinoma," *Cancer Causes and Control*, 12, 821–828.

Abnet, C. C., Lai, B., Qiao, Y. L., Vogt, S., Dong, Z. W., Taylor, P. R., Mark, S. D., and Dawsey, S. M. (2005), "Zinc Concentration in Esophageal Biopsies Measured by X-Ray Fluorescence and Cancer Risk," *Journal of the National Cancer Institute*, 97, 301–306.

Abnet, C. C., Qiao, Y. L., Dawsey, S. M., Buckman, D. W., Yang, C. S., Blot, W. J., Dong, Z. W., Taylor, P. R., and Mark, S. D. (2003), "Prospective Study of Serum Retinol, Beta Carotene, Cryptoxanthin, and Lutein/Zeaxanthin and Esophageal and Gastric Cancers," *Cancer Causes and Control*, 14, 645–655.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Blaser, M. J. (1999), "Hypothesis: The Changing Relationship of *Helicobacter pylori* and Humans: Implications for Health and Disease," *Journal of Infectious Diseases*, 179, 1523–1530.

Blot, W. J., and Li, J. Y. (1985), "Some Considerations in the Design of a Nutritional Intervention Trial in Linxian, People's Republic of China," *National Cancer Institute Monograph*, 69, 29–34.

Blot, W. J., Li, J. Y., Taylor, P. R., Guo, W., Dawsey, S., Wang, G. Q., Yang, C. S., Zheng, S. F., Gail, M., Yu, Y., Liu, B. Q., Tangera, J., Frauweni, J. F., Zhang, Y. H., and Li, B. (1993), "Nutrition Intervention Trials in Linxian, China: Supplementation With Specific Vitamin/Mineral Combinations, Cancer Incidence, and Disease-Specific Mortality in the General Population," *Journal of the National Cancer Institute*, 85, 1483–1492.

Borgan, O., Goldstein, L., and Langholz, B. (1995), "Methods for the Analysis of Sampled Cohort Data in the Cox Proportional Hazards Model," *The Annals of Statistics*, 23, 1749–1778.

Dawsey, S. M., Mark, S. D., Taylor, P. R., and Limburg, P. J. (2002), "Gastric Cancer and *H. pylori*," *Gut*, 51, 457–458.

Devesa, S. S., Blot, W. J., and Fraumeni, J. F. (1998), "Changing Patterns in the Incidence of Esophageal and Gastric Carcinoma in the United States," *Cancer*, 83, 2049–2053.

Helicobacter and Cancer Collaborative Group (2001), "Gastric Cancer and *Helicobacter pylori*: A Combined Analysis of 12 Case-Control Studies Nested Within Prospective Cohorts," *Gut*, 3, 347–353.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

International Agency for Research on Cancer (1994), *Schistosomes, Liver Flukes and Helicobacter pylori*, Lyon, France: Author.

Limburg, P. J., Wang, C. Q., Mark, S. D., Qiao, Y. L., Perez-Perez, G. I., Blaser, M. J., Taylor, P. R., Dong, Z. W., and Dawsey, S. M. (2001), "*Helicobacter pylori* Seropositivity: Association With Increased Gastric Cardia and Non-Cardia Cancer Risks in Linxian, China," *Journal of the National Cancer Institute*, 93, 226–233.

Mahabir, S., Abnet, C. C., Qiao, Y. L., Ratnasinghe, L. D., Dawsey, S., Dong, Z. W., Taylor, P. R., and Mark, S. D. (in press), "Polymorphisms of DNA Repair Genes XRCC1, XPD23, and APE5 and Risk of Stroke in Linxian, China," *Stroke*, in press.

Mark, S. D. (2003), "Nonparametric and Semiparametric Survival Estimation in Two-Stage (Nested) Cohort Studies," in *Proceedings of the American Statistical Association*, pp. 2675–2691.

Mark, S. D., and Katki, H. (2001), "Influence Function–Based Variance Estimation and Missing Data Issues in Case-Cohort Studies," *Lifetime Data Analysis*, 7, 329–342.

Mark, S. D., Selhub, J., Qiao, Y. L., Buckman, D., Dawsey, S. M., Blot, W. J., Dong, Z. W., and Taylor, P. R. (2001), "Serum Cysteine and Riboflavin Are Inversely Related to Incident Esophageal and Gastric Cardia Cancers," in *Proceedings, American Association of Cancer Research*, New Orleans, LA.

Mark, S. D., Qiao, Y. L., Dawsey, S. M., Katki, H., Gunter, E. W., Yan-Ping, W., Fraumeni, J. F., Blot, W. J., Dong, Z. W., and Taylor, P. R. (2000), "Higher Serum Selenium Is Associated With Lower Esophageal and Gastric Cardia Cancer Rates," *Journal of the National Cancer Institute*, 92, 1753–1763.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

Pugh, M. G. (1993), "Inference in the Cox Proportional Hazards Model With Missing Covariate Data," thesis, Harvard School of Public Health, Boston, MA.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

Roth, M. J., Abnet, C. C., Johnson, L. L., Mark, S. D., Dong, Z. W., Taylor, P. R., Dawsey, A. M., and Qiao, Y. L. (2004), "Polymorphic Variation of CYP1A1 Is Associated With the Risk of Gastric Cardia Cancer: A Prospective Case-Cohort Study of Phase I and Phase II Cytochrome P-450 1A1 and GST Enzymes," *Cancer Causes and Control*, 15, 1077–1083.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Samet, J. M., and Munoz, A. (1998), "Evolution of the Cohort Study," *Epidemiologic Reviews*, 20, 1–14.

Savage, S. A., Abnet, C. C., Haque, K., Mark, S. D., Qiao, Y. L., Dong, Z. W., Dawsey, S. M., Taylor, P. R., and Chanock, S. J. (2004a), "Polymorphisms in Interleukin 2 and Interleukin 10 Are Not Associated With Gastric Cardia or Esophageal Cancer in a High-Risk Chinese Population," *Cancer Epidemiology Biomarkers and Prevention*, 13, 1547–1549.

Savage, S. A., Abnet, C. C., Mark, S. D., Qiao, Y. L., Dong, Z. W., Dawsey, S. M., Taylor, P. R., and Chanock, S. J. (2004b), "Variants of the IL8 and IL8RB Genes and Risk for Gastric Cardia Adenocarcinoma and Esophageal Squamous Cell," *Cancer Epidemiology Biomarkers and Prevention*, 13, 2251–2257.

Self, S. G., and Prentice, R. L. (1988), "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies," *The Annals of Statistics*, 16, 64–81.

Stolzenberg-Solomon, R., Abnet, C. C., Ratnasinghe, L., Qiao, Y. L., Dawsey, S. M., Dong, Z. W., Taylor, P. R., and Mark, S. D. (2003), "Esophageal and Gastric Cardia Cancer Risks and MTRR A66G and MTHFR C677T and A1298C Polymorphisms in Linxian, China," *Cancer Epidemiology Biomarkers and Prevention*, 12, 1222–1226.

Taylor, P. R., Qiao, Y. L., Abnet, C. C., Dawsey, S. M., Yang, C. S., Gunter, E. W., Blot, W. J., Dong, Z. W., and Mark, S. D. (2003), "Prospective Study of Serum Vitamin E Levels and Esophageal and Gastric Cancers," *Journal of the National Cancer Institute*, 95, 1414–1416.